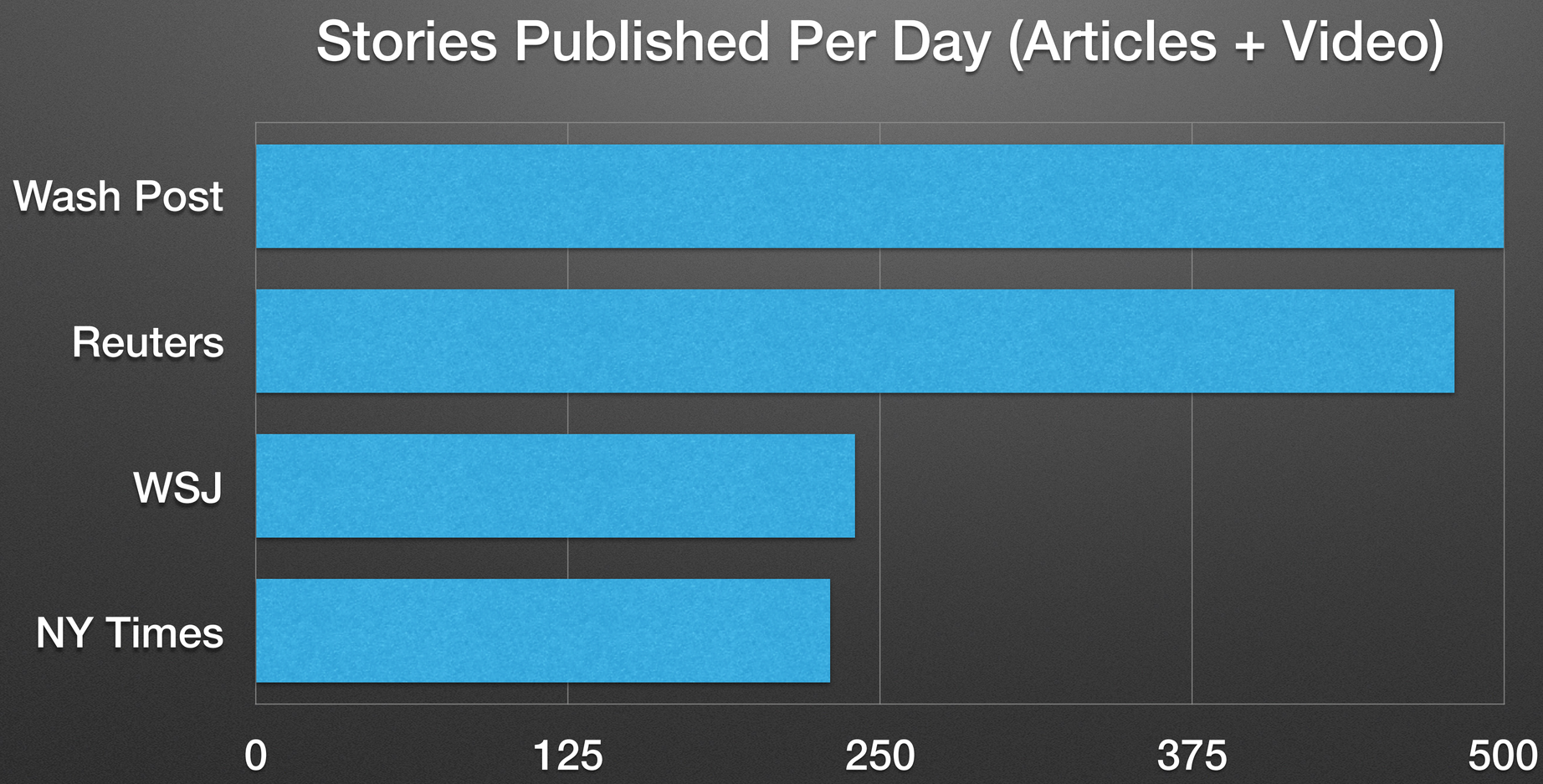
Auto-Gist the News

With Topic Modeling and Extractive Text Summarization Methods

March 2, 2018

Background


Stories Published Per Day (Articles + Video)
Wash Post
Reuters
WSJ
NY Times
0
125
250
375
500

The Data

30,000

Reuters News Articles
(January 1, 2018 ~ Present)

- Article URLs pulled using News API (contains links to articles from over 5,000 news sources and blogs)
- Scrapy / BeautifulSoup for scraping content

Topic Modeling

- TF-IDF to reduce weight of terms frequent across documents
- Non-Negative Matrix Factorization (NMF) to extract document topics
- **30 topics total**

Industry-Specific

AIRCRAFT	*boeing, airbus, embraer, bombardier, jets*
AUTOMOTIVE	*gm, vehicles, electric, ford, cars*
BUSINESS	*percent, billion, quarter, company, revenue*
FINANCIAL	*bank, banks, billion, financial, funds*

Country / Region-Specific (Political)

IRAN	*iran, iranian, nuclear, sanctions, tehran*
ISRAEL / PALESTINE	*israel, israeli, jerusalem, palestinian*
NORTH KOREA	*north, korea, korean, south, kim, nuclear*
SAUDI ARABIA	*saudi, arabia, aramco, prince, yemen*
TURKEY / SYRIA	*turkey, syria, syrian, turkish, ypg*

Text Summarization

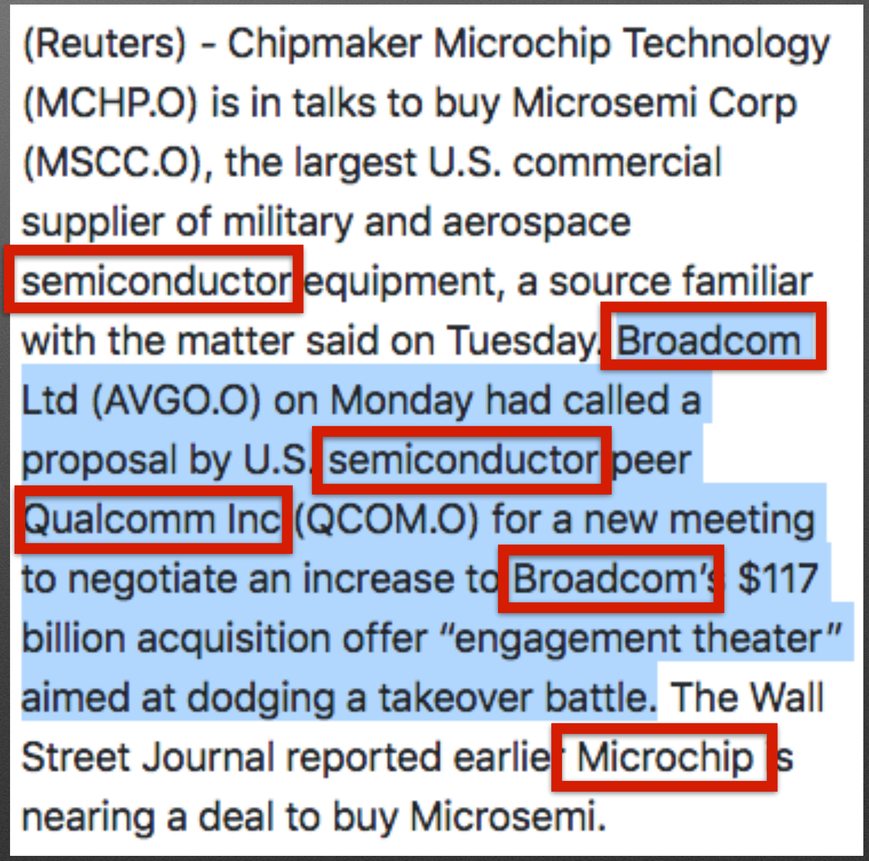
- 7 Sentence Extraction Algorithms Tested:
  - Luhn
  - Edmundson
  - Lexical Rank
  - Text Rank
  - Sum Basic
  - Latent Semantic Analysis
  - Kulback-Lieber

Luhn Summarizer

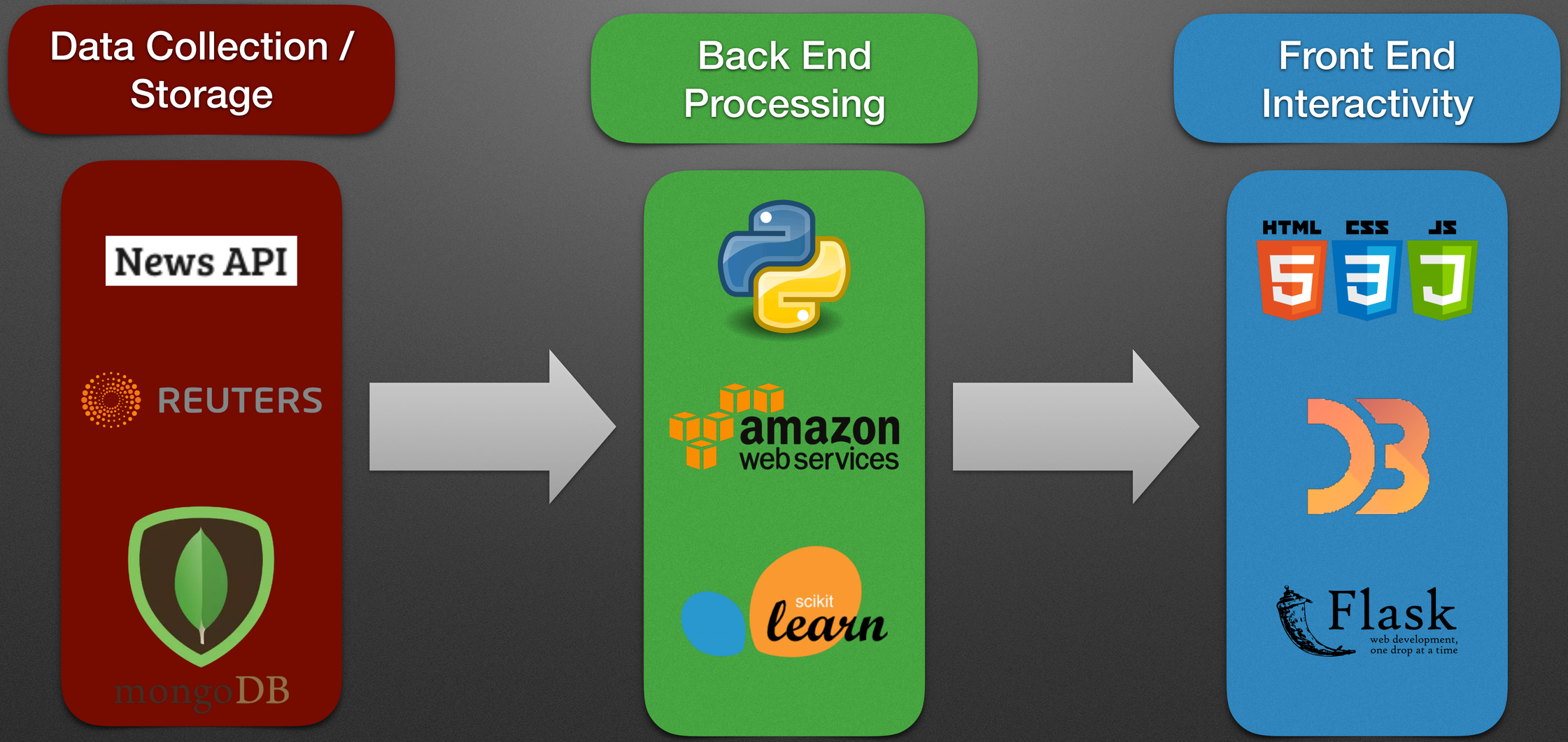
- Term frequency determines sentence importance
- TF-IDF for word weighting in document
- Stop word filtering
- Cluster of frequent words indicates good sentence

Edmundson Summarizer

- Four weighted features for sentence importance:
  - Cue words (e.g. “Significant”, “Greatest”, “Impossible”, “Hardly”)
  - Title & heading words
  - Key word frequency (related to topic)
  - Sentence location



Summarization In Practice: YRNWS



YRNWS

It's Your News, But Shorter

Date Range

February 24, 2018 - February 24, 2018

Blog / Website

- Reuter's (www.reuters.com/articles)

Topics of Interest

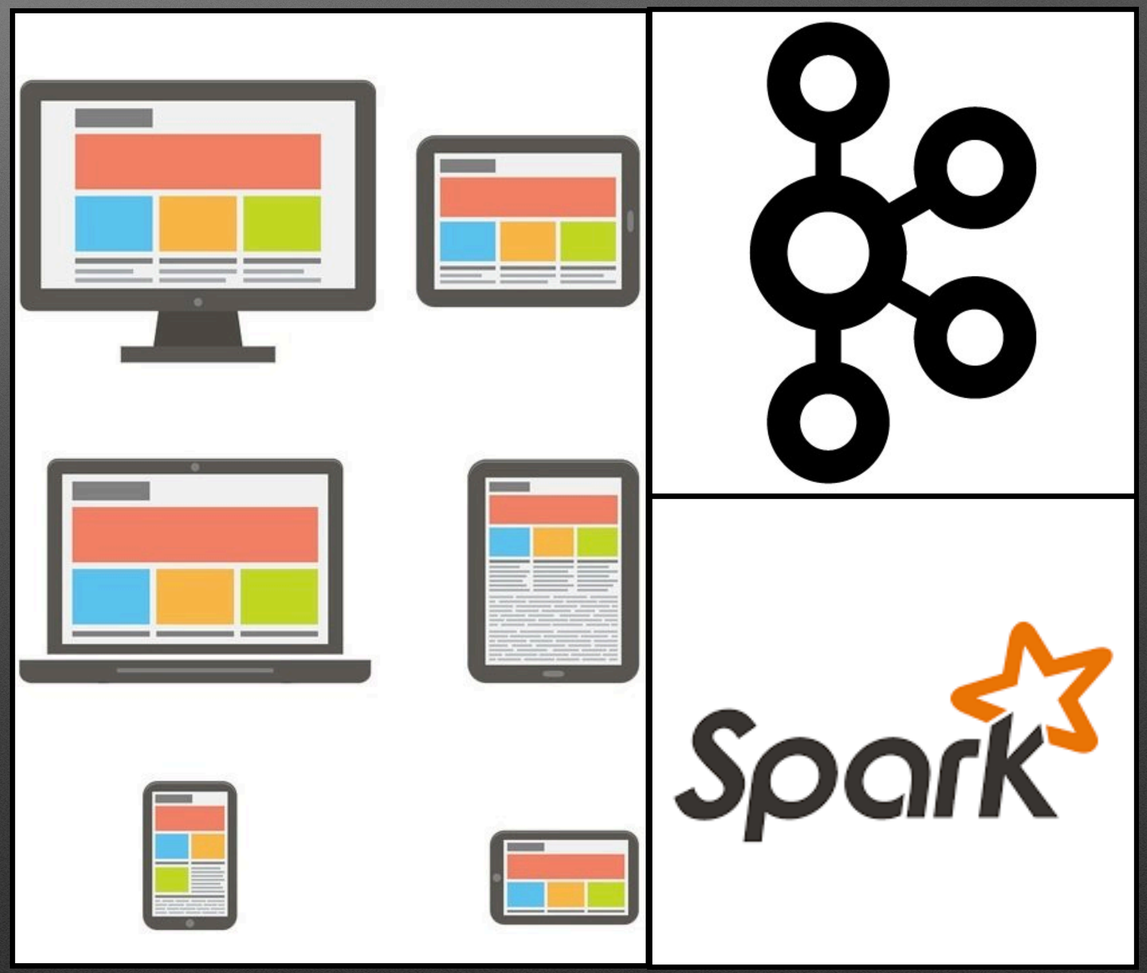
Summary Length

Auto (Currently 3 Sentences)

Generate Report

Next Steps

- Enhanced Data Storage and Streaming
- Web Deployment
- Addition of Other News Sources



Thank You

Topic Modeling

Reuters News

AFGHANISTAN / PAKISTAN	*pakistan, afghanistan, taliban, islamabad, afghan*	IRAN	*iran, iranian, nuclear, sanctions, tehran*
AFRICA	*zuma, anc, ramaphosa, africa, south*	ISRAEL / PALESTINE	*israel, israeli, jerusalem, palestinian, palestinians*
AIRCRAFT	*boeing, airbus, embraer, bombardier, jets*	NORTH AMERICA	*canada, canadian, nafta, trade, mexico*
ASIA	*china, chinese, beijing, trade, kong*	NORTH KOREA	*north, korea, korean, south, kim, nuclear*
AUTOMOTIVE	*gm, vehicles, electric, ford, cars*	OLYMPICS (HIGHLIGHTS)	*gold, olympic, medal, team, pyeongchang*
BUSINESS	*percent, billion, quarter, company, revenue*	OLYMPICS (SCANDALS)	*doping, athletes, ioc, russian, russia*
CRIME / COURT CASES	*court, case, supreme, justice, law*	SAUDI ARABIA	*saudi, arabia, aramco, prince, yemen*
ECONOMY	*inflation, percent, growth, rate, economy*	SECURITY / TERRORISM	*police, people, killed, city, attack*
ENVIRONMENTAL	*oil, crude, bpd, production, opec*	SOUTH AMERICA	*maduro, venezuela, colombia, opposition*
EU (BREXIT)	*eu, britain, brexit, european, london*	SPORTS	*league, game, season, club, team*
EUROPE (POLITICS)	*party, government, minister, election, parliament*	STOCK MARKET	*percent, index, stocks, points, dollar*
FINANCIAL	*bank, banks, billion, financial, funds*	TECHNOLOGY	*qualcomm, broadcom, apple, nxp, chips*
GERMANY (POLITICS)	*spd, merkel, coalition, germany, conservatives*	TENNIS	*match, open, australian, slam, federer*
HEALTH / MEDICINE	*study, health, patients, women, drug*	TURKEY / SYRIA	*turkey, syria, syrian, turkish, ypg*
IMMIGRATION	*myanmar, rohingya, rakhine, bangladesh, refugees*	U.S. POLITICS	*trump, house, republican, white, democrats*